Pilsen Soils OU2 Site
(Superfund Removal Program)
Simple Linear Regression and Diagnostics Results for Lead
(2016-2017 Sampling Event)

Prepared by
John Canar
FIELDS Group, US EPA, Region V

17 July 2017

Summary

The regression of Lead XRF and laboratory values was statistically significant. Using the below regression equation, a Lead XRF value of 369ppm corresponds to a Laboratory value of 400ppm. However, given the uncertainty associated with XRF and laboratory data and to use the uncertainty found in the below best-fit regression equation for future XRF sampling results, employing a prediction limit will ensure with a known confidence that a Lead XRF value is indeed below 400ppm. Using the upper prediction limit, a Lead XRF value of 290ppm has only a 5% chance of being a 400ppm had that sample been sent to the laboratory.

This statistical evaluation was completed to allow the USEPA On-Scene Coordinator (or their Representative) to conduct field tests using an XRF (SN 510349) to determine if the residential cleanup goal of 400ppm has been met prior to reaching the target depth of 2-feet in gardens.

The FIELDS Group would recommend the following decision criteria with the XRF (SN 510349):

- if the median from five XRF readings from a sample at the base of the excavation is <= 290ppm, the excavation pit soil can be considered clean and be filled;
- if the median from five XRF readings from a sample at the base of the excavation is >= 369ppm, the excavation pit surface needs further remediation;
- if the median from five XRF readings from a sample at the base of the excavation is > 290ppm and < 369ppm, there are two options:
  - send the sample to the lab to determine if indeed the sample has a value < 400ppm, or
  - dig an addition 3 inches, collect a new composite sample, re-shoot with the XRF. Continue this process until the median of the five Lead XRF readings are <= 290ppm.

Introduction

Simple linear regression and regression diagnostics were used to find the best fitting linear relationship between XRF measurements of Lead levels in soil samples and their corresponding laboratory measurements using the SAS® software. This relationship is quantified into a model (equation) of XRF measurements of Lead and their corresponding laboratory measurements. The statistical methods employed were drawn from SAS® literature and three regression texts: Statistical Methods in Water Resources, 1992; and Applied Regression Analysis and Other Multivariate Methods, 1978 and 1988. (See "References" section for a complete list of regression resources.) The data set used for this analysis was provided by TetraTech, the USEPA contractor for the Pilsen Soils Superfund site. The data are from samples collected in 2016 and 2017 in OU2. This site is under the direction of Ramon Mendoza, USEPA OSC.

The steps used to perform simple linear regression were:

1. Plot the data;
2. Compute the least squares regression statistics;
3. Examine adherence to the assumptions of regression using residual plots; and
4. Employ regression diagnostics (Helsel and Hirsch, 1992).

Data and data handling

A total of 36 soil samples with corresponding XRF values that met the following criteria were used for regression: represented a 6-inch interval, were in a residential yard, and were not in a vacant lot. (These 36 samples are shown in Table 1.) Hence vacant lot and garden samples were excluded. The former were excluded as they are often disturbed properties, e.g., the site of a former home that was removed, and hence, soil was disturbed in the process. The latter were excluded as they are often 0-12 inches and or represent very small areas.

The data files were provided by TetraTech, the USEPA contractor. The original files are "Table2_Final Sample Results_Pilsen OU2_asof061317.xlsx" and Pilsen_OU2_XRFData.csv". Each sample had five XRF readings for Lead except for one sample that accidentally had four readings. None of the laboratory or XRF data had limit of detection values. The median value of the five, or four in one case, XRF readings was used for regression analysis.

An Olympus, formerly Innov-X, Delta 4000 XRF analyzer was used to analyze each of the above 36 soil samples. The serial number for this XRF is 510349. The regression equation found below is specific to this XRF device. Hence, substitution of a different XRF device may lead to a different regression relationship between XRF results and laboratory results.

Results

There was a statistically significant linear regression relationship between XRF Lead values and their corresponding Laboratory values (results not shown). However, regression diagnostics

found that some of the assumptions of regression were violated. These violations included extreme residuals, heteroscedasticity, and non-normality of the residuals (see Figures 1 and 2). (The null hypothesis of each of the four tests in Figure 2 is that the residuals are from a normal distribution. If using an alpha value of 0.05, one would reject the null hypothesis for all four tests.) To overcome these violations, two observations with Studentized residual values greater than 2.5, a value used as a rule of thumb for extreme values, were removed from the data set. The new data set was regressed and the linear regression was significant (results not shown). However, regression diagnostics found that some of the assumptions of regression were violated. These violations included an extreme residual, heteroscedasticity, and non-normality of the residuals (results not shown). To overcome these violations, another observation with a Studentized residual value greater than 2.5 was removed from the data set. The new data set was regressed and the linear regression was significant (results not shown). However, the residuals still exhibited a fair amount of heteroscedasticity. A data transformation would likely overcome this violation of the statistical assumptions of regression. Hence, the natural log of the XRF Lead values and their corresponding Laboratory values were taken.

There was a statistically significant linear regression relationship between the natural log of XRF Lead values and their corresponding natural log of Laboratory values (results not shown). The heteroscedasticity is now much less apparent. However, regression diagnostics found an extreme residual. The one observation with a Studentized residual value greater than 2.5was removed from the data set. The new data set was regressed and the linear regression was significant (results not shown). An additional extreme residual occurred. That observation was removed from the data set and the new data set was regressed. The regression results were significant and the assumptions of regression were met. Figure 3 shows the statistically significant linear regression relationship between the natural log of XRF Lead values and their corresponding natural log of Laboratory values. Figures 4 and 5 demonstrate that the assumptions of regression were met. Figure 4 shows that the residuals were homoscedastic and none of the Studentized residuals were greater than 2.5. Figure 5 shows that the residuals were normally distributed. Normality of residuals is required in order to test the hypothesis that "the slope coefficient (β_1) is significantly different from zero" (Helsel and Hirsch, 1992). In other words, in order to demonstrate a linear relationship between the two variables, XRF and Laboratory, the slope coefficient must be significant. A visualization of the linear relationship between the natural log of Lead XRF and Laboratory values in soil is shown in Figure 6.

The parameters of the best linear fit equation for the relationship of the natural log of Lead XRF and Laboratory values in soil are:

Adjusted LN Lead = -0.39534 + (1.08023)*(LN XRF Lead value)

However, as this equation is in natural log space, the antilog of the adjusted Lead value must be taken. For example, for an XRF Lead reading of 400ppm (5.99ppm in natural log space), the Adjusted LN XRF Lead value is 6.077ppm. The antilog of this value is 436ppm. Hence, an XRF Lead reading in soil of 400 ppm is equivalent to an adjusted XRF Lead value of 436ppm in soil. (436ppm is the estimated laboratory value.) An adjusted XRF Lead value of 400ppm is equivalent to an XRF Lead reading of 369ppm.

Given the uncertainty associated with XRF and laboratory data and to use the uncertainty found in the above regression equation for future XRF sampling results, employing a prediction limit will ensure with a known confidence that a Lead XRF value is indeed below 400ppm. The UPL (upper prediction limit) provides a very good estimate of what the lowest XRF Lead value can be in order to be 95% confident that in future sampling that value would be less than 400ppm had that sample been sent for Laboratory analysis. The current UPL95 value for the median Lead XRF is 290ppm. Hence, we'd expect only 5% of samples with an XRF value less than 290ppm to be 400ppm or above in the lab. (95% and 5% are derived by splitting the 10% for both sides of the UPL and LPL lines, i.e., the 90% prediction limits as seen in Figure 7.)

Figure 7 shows the confidence limits (blue lines) and prediction limits (red lines), as well as the gray zone for the Lead XRF and Laboratory values regression equation. The figure visualizes the UPL (upper-most red line) and its relationship to the confidence limits (blue lines) as well as the best-fit regression line (displayed as a series of green dots). The range of XRF values between the UPL and the predicted Lead value (the adjusted XRF value) is the gray zone. This is the zone in which a project manager may choose to send samples to the laboratory to determine if the sample is below 400ppm. For the regression results, this zone has XRF Lead values between 290ppm and 369ppm, where 369ppm is the median Lead XRF value that adjusts to a 400ppm via regression.

Using Figure 7 as a visual guide, the FIELDS Group would recommend the following decision criteria with the XRF (SN 510349):

- if the median from five XRF readings from a sample at the base of the excavation is <= 290ppm, the excavation pit soil can be considered clean and be filled;
- if the median from five XRF readings from a sample at the base of the excavation is >= 369ppm, the excavation pit surface needs further remediation;
- if the median from five XRF readings from a sample at the base of the excavation is > 290ppm and < 369ppm, there are two options:
    - send the sample to the lab to determine if indeed the sample has a value < 400ppm, or
    - dig an addition 3 inches, collect a new composite sample, re-shoot with the XRF. Continue this process until the median of the five Lead XRF readings are <= 290ppm.

References:

Chen, X., Ender, P., Mitchell, M. and Wells, C. (2003). Regression with SAS, from http://www.ats.ucla.edu/stat/sas/webbooks/reg/default.htm

Helsel, D.R. and Hirsch R.M., Statistical Methods in Water Resources, Elsevier, Amsterdam, 1992.

Kleinbaum, D.G. and Kupper, L.L., Applied Regression Analysis and Other Multivariate Methods, Duxbury Press, Boston, Massachusetts, 1978.

Kleinbaum, D.G., Kupper, L.L., and Muller, K.E., Applied Regression Analysis and Other Multivariate Methods, Second Edition. PWS-Kent Publishing Company, Boston, Massachusetts, 1988.

SAS Institute Inc., SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999. (Chapter 55, The REG Procedure)

SAS Institute Inc., SAS ® System for Regression, Second Edition, Cary, NC: SAS Institute Inc., 1991. 210pp.

Contact:

Please contact John Canar (canar.john@epa.gov or 312.886.6182) about this document.

SampleID	Lab Lead ppm	XRF Lead ppm
S-160407-GW-025-ES-D	167	79
S-160407-GW-025-ES	91.4	100
S-160412-GW-047-ES	174	179
S-160406-GW-023-ES	163	187
S-160407-GW-033-ES	425	377
351-01-032416	486	382
S-160512-GW-052-ES	543	456
S-160405-GW-017-ES	553	481
S-160407-GW-032-ES	734	548
S-160411-GW-036-ES	737	557
S-160411-GW-034-ES	672	573
186-01-032316	568	623
S-160909-AK-060-ES	761	781
S-160411-GW-035-ES	778	782
186-02-032316	753	783
S-160407-GW-028-ES	1410	809
S-160406-GW-018-ES	1070	937
S-120716-WP-061-ES	986	1031
S-160406-GW-019-ES	1560	1036
S-160405-GW-012-ES	1260	1059
S-160404-GW-001-ES	1470	1183
S-160411-GW-041-ES	1690	1256
S-160407-GW-029-ES	1680	1418
S-160406-GW-020-ES	1780	1530
S-051117-GW-79-ES	2100	1659
S-160512-GW-055-ES	2240	1668
S-160404-GW-009-ES	2580	1758
S-051117-GW-78-ES	2340	1848
S-160404-GW-005-ES	3340	2007
S-160512-GW-057-ES	2090	2058
S-160512-GW-049-ES	2630	2084
S-160512-GW-048-ES	3120	2123
S-041317-WP-66-ES	2910	2201
S-160404-GW-007-ES	2880	3074
S-160407-GW-030-ES	4050	3390
S-160404-GW-003-ES	4400	3902

Table 1: Laboratory and XRF sample IDs and Lead values. Note: the XRF_Lead_ppm values are medians

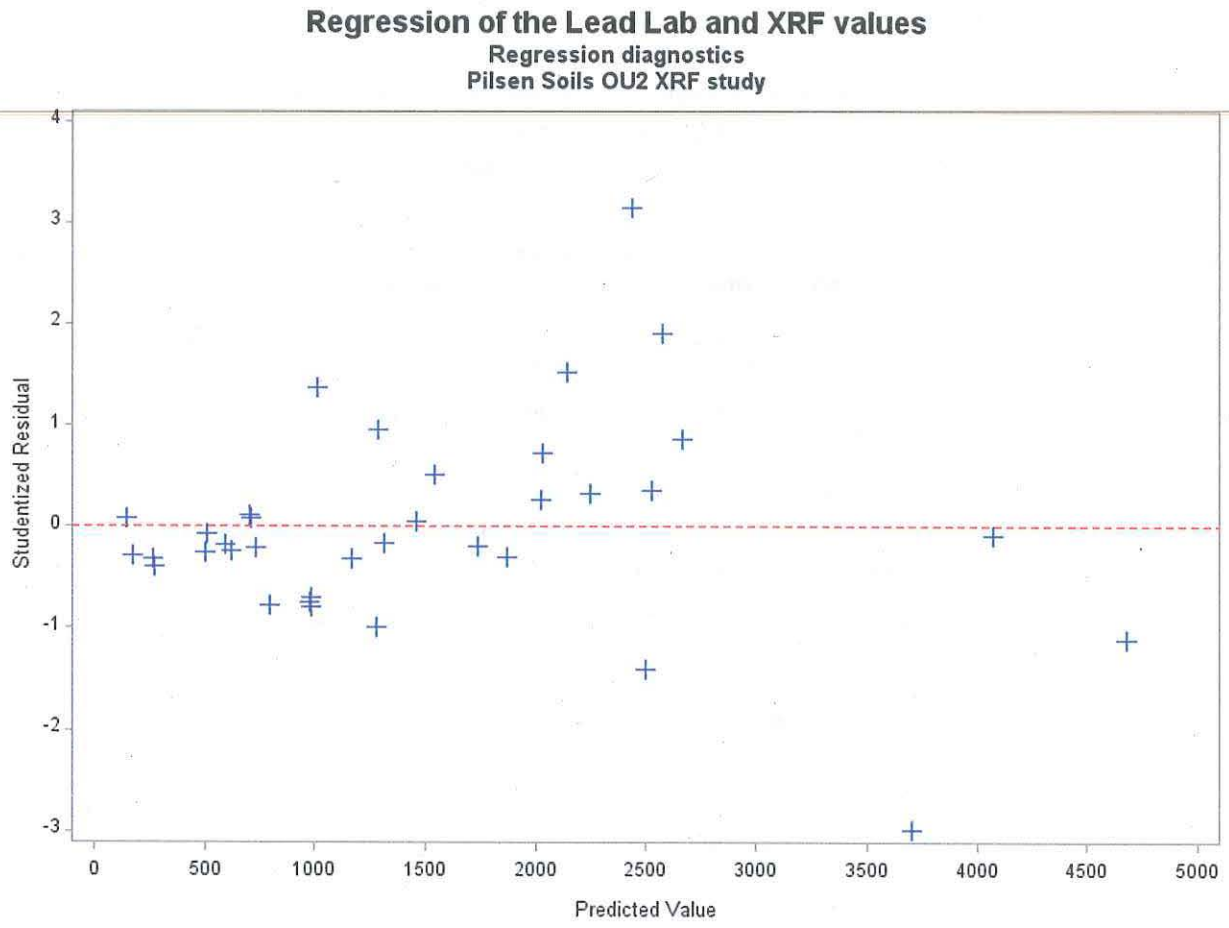

Figure 1: Residual plot from the SAS software for the Lead XRF and Laboratory values

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.922958	Pr < W	0.0154
Kolmogorov-Smirnov	D	0.151706	Pr > D	0.0353
Cramer-von Mises	W-Sq	0.201696	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.098874	Pr > A-Sq	0.0064

Figure 2: Tests of Normality from the SAS software for residuals from the Lead XRF and Laboratory values

Regression of the Natural Log of Lead Lab and XRF values
Regression diagnostics
Pilsen Soils OU2 XRF study

The REG Procedure
Model: MODEL1
Dependent Variable: LN_lab Natural Log of Lab Lead (ppm)

Number of Observations Read	34
Number of Observations Used	34

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.49535	28.49535	1289.87	<.0001
Error	32	0.70693	0.02209		
Corrected Total	33	29.20228			

Root MSE	0.14863	R-Square	0.9758
Dependent Mean	7.02583	Adj R-Sq	0.9750
Coeff Var	2.11552		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.39534	0.20820	-1.90	0.0666
LN_XRF	Natural Log of XRF Lead (ppm)	1	1.08023	0.03008	35.91	<.0001

Figure 3: Simple linear regression output from the SAS software for the natural log of the Lead XRF and Laboratory values

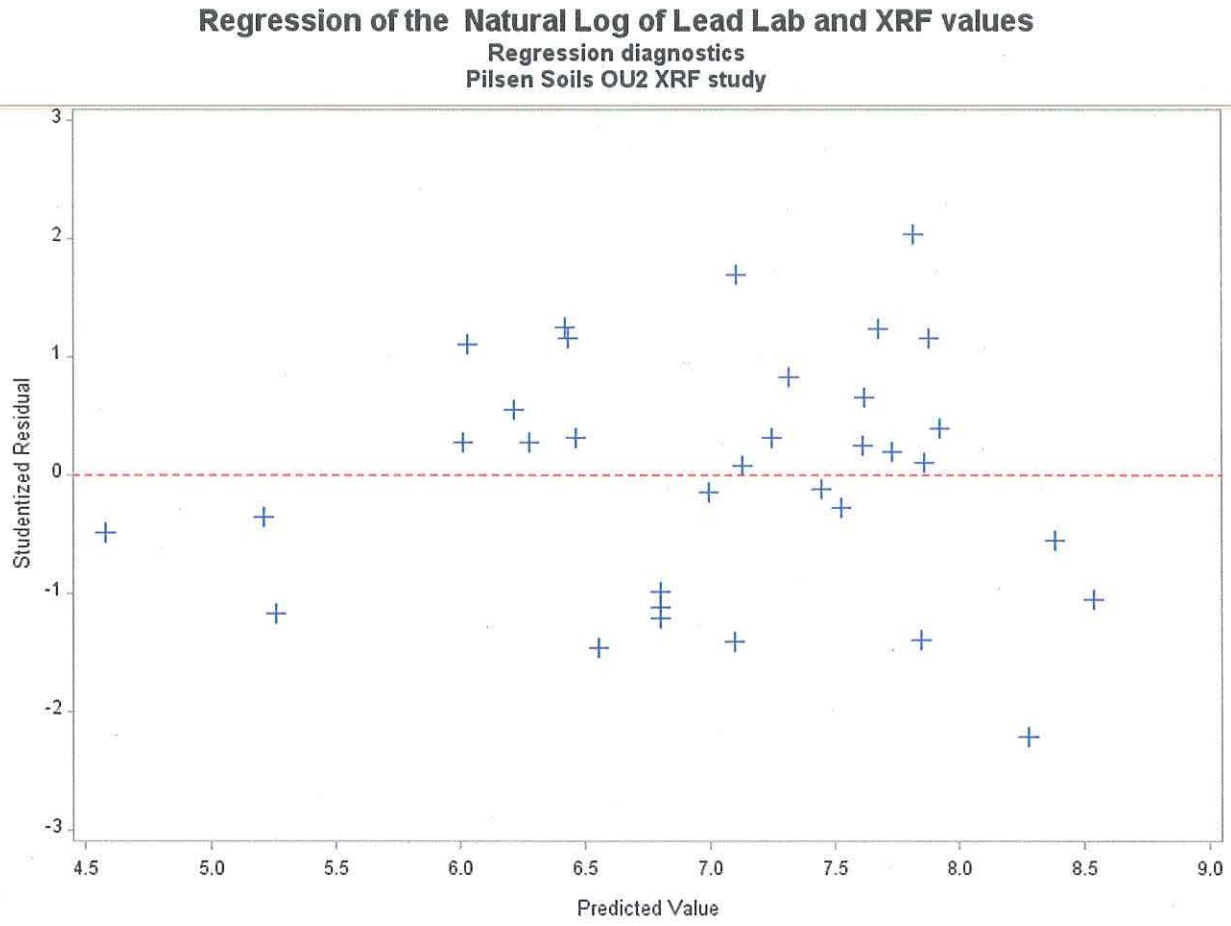

Figure 4: Residual plot from the SAS software for the natural log of Lead XRF and Laboratory values

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.978577	Pr < W	0.7274
Kolmogorov-Smirnov	D	0.10259	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.049955	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.313263	Pr > A-Sq	>0.2500

Figure 5: Tests of Normality from the SAS software for residuals from the natural log of Lead XRF and Laboratory values

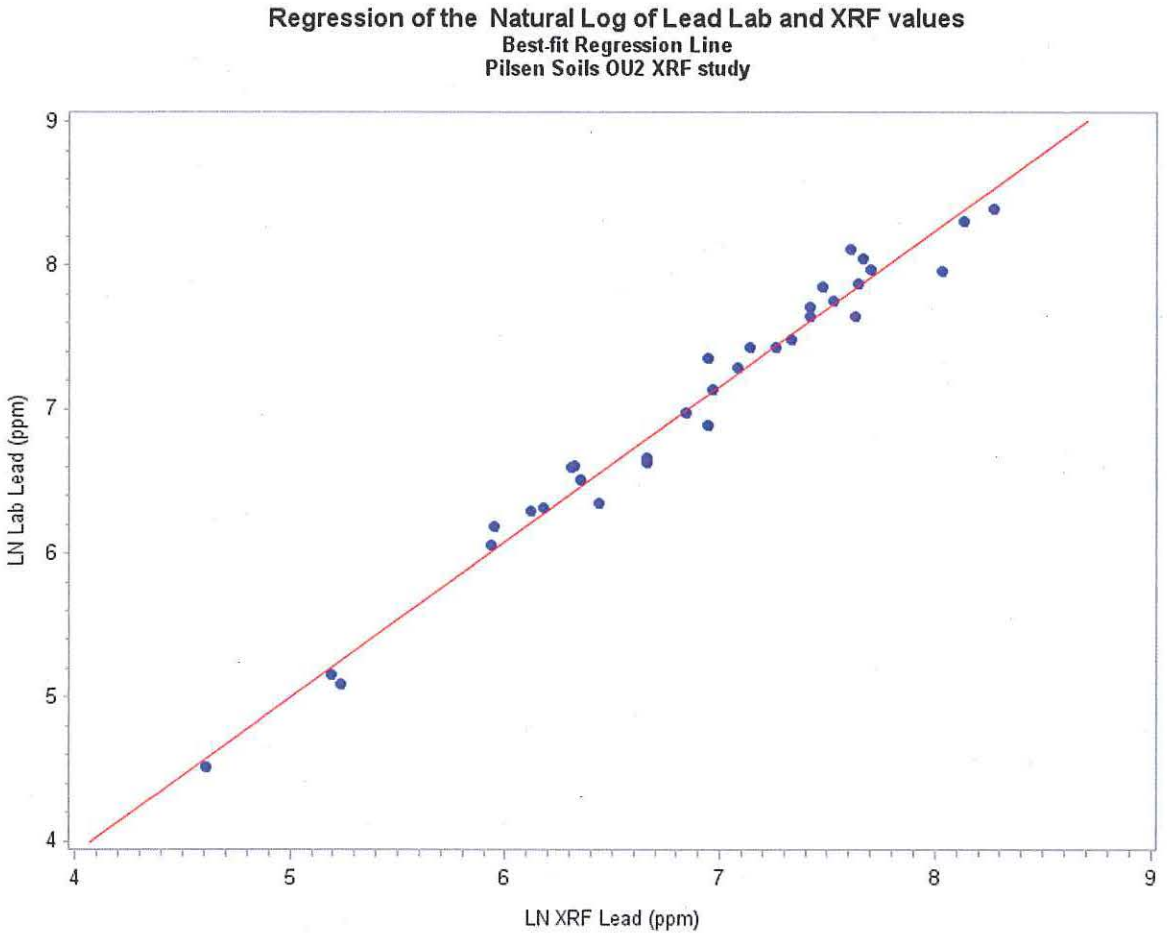

Figure 6: Best-fit linear regression line from the SAS software for the natural log of the Lead XRF and Laboratory values

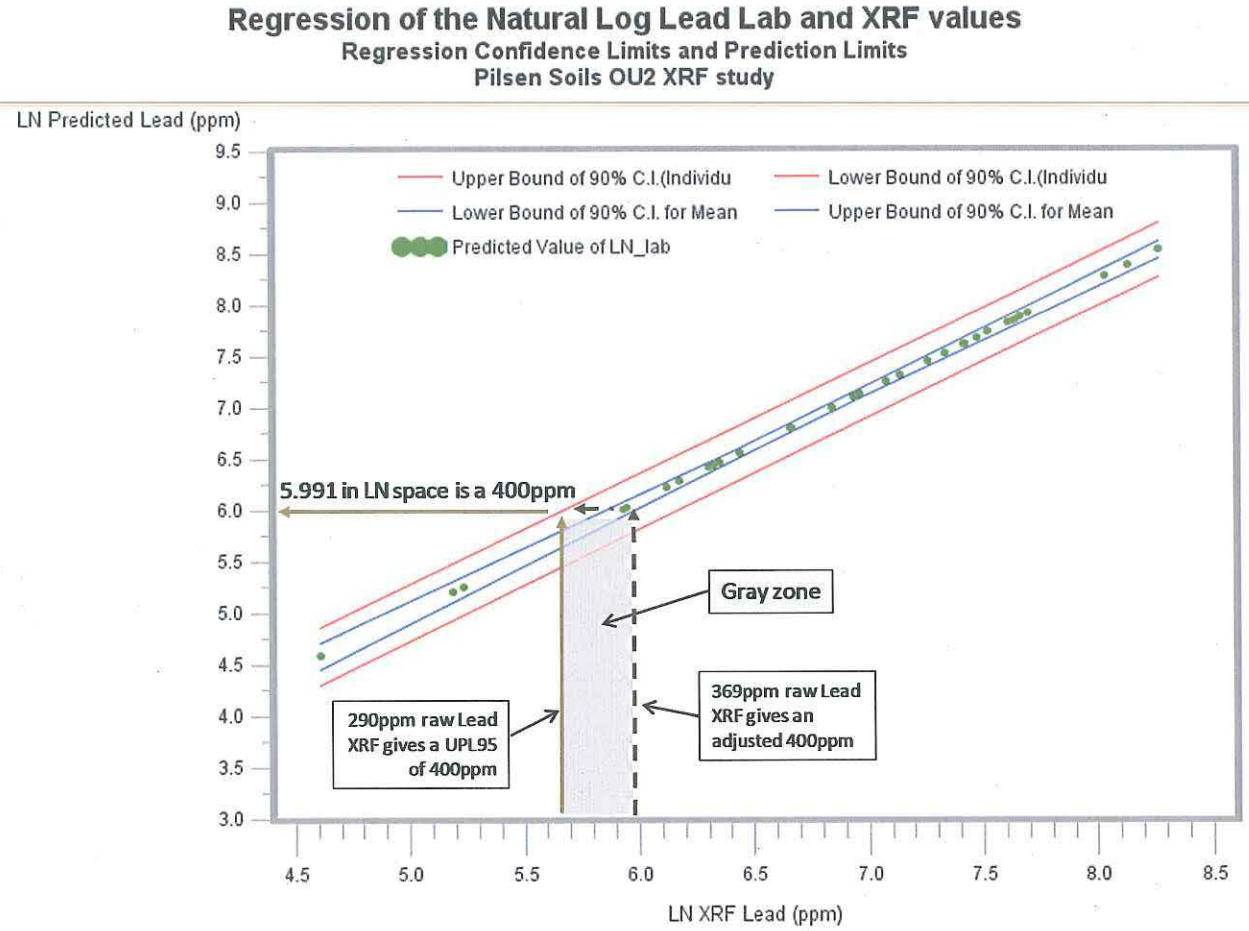

Figure 7: Best-fit linear regression line from the SAS software for the natural log of the Lead XRF and Laboratory values, including confidence limits (in blue) and prediction limits (in red)